J Value Inquiry (2011) 45:85–90
DOI 10.1007/s10790-011-9265-1

BOOK REVIEW

Herbert Gintis, The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences

Princeton, New Jersey: Princeton University Press, 2009. 304 pp. ISBN 978-0-691-14052-0, $37.50 (Hb)

John Thrasher

Published online: 17 March 2011

Humans are social animals. We have been shaped both by biological evolution and the logic of strategic interaction that organizes our social world. In the *Bounds of Reason*, Herbert Gintis argues that the techniques of game theory, combined with evidence from experimental data on human psychology and behavior will provide a unified vision of the social and behavioral sciences. This grand reconciliation project aims to create a common science of human behavior that would unify economics, psychology, sociology, and anthropology into a common field. This project is promising for philosophers as well as social scientists and should not only interest philosophers working in decision theory, epistemology, practical reasoning, and moral psychology, but also those concerned with social and political philosophy.

The *Bounds of Reason* begins with the claim that, "game theory is central to understanding the dynamics of life forms in general, and humans in particular" (p. xiii). Game theory is a method of modeling strategic interactions between agents and their environments. It is a powerful tool, a magical tool, as Gintis claims, to understand and model the logic of social interaction. Social and behavioral sciences that abandon game theory are, according to Gintis, seriously handicapped. This is not to say that Gintis is merely advocating that game theory, as it is currently practiced, will be a magic bullet. He is critical of several important aspects of modern game theory and much of the *Bounds of Reason* is an attempt to defend a version of behavioral game theory that incorporates behavioral data from experimental and behavioral economics as well as psychology. Modern game theory is, according to Gintis, often in conflict with the behavior of actual decision makers. Gintis argues that the experimental evidence has increasingly shown that "rational agents just do not behave the way classical game theory predicts" (p. xvii). They do not play mixed strategies, use backwards induction, or always play Nash equilibriums—all key concepts in modern game theory. This experimental data,

J. Thrasher (✉)
University of Arizona, Tucson, AZ, USA
e-mail: jthrashe@email.arizona.edu

however, has not been used to alter or expand the assumptions used by game theorists. The problem, according to Gintis, is that the game theorists are not talking to the experimentalists. Furthermore, the experimentalists have also failed to develop a plausible model of individual rationality that can explain their own data. Unifying the behavioral sciences requires a model of individual and strategic rationality that can explain and organize a systematic investigation into our social life. What we need is a plausible and usable model of rationality.

According to Gintis, the best model of individual rationality available is the "beliefs, preferences, constraints" model that sees the rational agent as "an individual with consistent preferences" (p. 1). Consistency here is nothing more than transitivity of preferences at a particular time. Transitivity of preference is axiomatic, but it is also not empirically unmotivated. Gintis argues that organisms with consistent preferences will tend to fare better than those with inconsistent preferences; hence there is reason to believe that consistency will have been adaptive. This model, simple though it may seem, incorporates many of the psychological heuristics and potential biases that have become so popular with behavioral economists and psychologists. This literature has led some to believe that we should reject the basic beliefs, preferences, constraints model of rationality because it does not explain commonly observed behavior. Gintis, however, argues that, "human beings did not evolve facing general decision-theoretic problems" (p. 29). The fact that we are often not good at solving these problems should not lead us to think that we inherently irrational. The analytical strength of the beliefs, preferences, constraints model combined with the well-established experimental data allows us to model human behavior in a realistic and consistent way.

Decision theory, however, only deals with a small range of human reasoning. Individual rational choice under uncertainty must be expanded to include a theory of the rationality of dynamic interaction with other agents. The theoretical foundation of the science of strategic choice—game theory—is the Nash Equilibrium. A Nash equilibrium is a best response to the best response of the other players in a game. Simply put, in a Nash equilibrium, no player gains from unilaterally changing their strategy. For any game that meets some minimal standards, there is at least one Nash equilibrium. The problem, however, is that in any given game there are often more than one Nash equilibrium. Because of this, many equilibrium refinement techniques have been proposed to reduce the number of equilibrium strategies. Certain assumptions about the rationality of agents in the game can make this elimination easier.

One of the main arguments in the *Bounds of Reason* is that the Nash equilibrium should be replaced by the less well-known idea of a correlated equilibrium. For the Nash equilibrium to be a plausible equilibrium for actual human agents, we need to make certain assumptions about the common knowledge shared between persons in the game. Game theorists and epistemologists commonly move from the fact that agents are mutually rational to the idea that they have Common Knowledge of Rationality. This later condition is much stronger and does not follow from the fact that each agent is individually rational. Common Knowledge of Rationality holds when, "each player is rational, each knows the others are rational, each knows the others know the others are rational and so on" (p. 92). But, as Gintis argues, "there is no justification for such reasoning"(p. 93). Common Knowledge of Rationality is not a plausible model of our epistemic situation, nor is it necessary to explain human interaction. The Bayesian model of reasoning that Gintis prefers does not assume Common Knowledge of Rationality. Which is not to say that Common Knowledge of Rationality can never obtain, it just cannot be assumed as a straightforward implication of individual rationality. Rather, Common Knowledge of Rationality can be an event within the game. Agent may be able to discern something like common knowledge from the move of another agent, but that knowledge is itself the result of inferences from action, not a starting point.

This argument has important implications for how we model rational human interaction due to the relationship between Common Knowledge of Rationality and backwards induction. Backwards induction involves eliminating weakly and strongly dominated strategies from a player's set of strategies. Strongly dominated strategies are not Nash equilibriums, but weakly dominated strategies may be, though there is another strategy that is at least as good or better than the weakly dominated strategy. Backwards induction with Common Knowledge is useful because it allows the elimination of Nash equilibriums that are weakly dominated. Imagine a game in extensive form as a decision tree. Using backwards induction, one starts from the final decision node of the game and eliminates each option that is worse than any other option available to the player at that node. Branches of the game tree that have not been eliminated represent the best moves available at each decision node. But if Common Knowledge of Rationality is an event not a premise, we notice that there is nothing inherently irrational about a player using a weakly dominated strategy in a given game. Players will use the moves of the other players as data to determine the rationality of the player. Instead of backwards induction, players will use "forward induction" to make bets about what other players will do in the future given the moves they have made in the past. Players who play weakly dominated strategies will not necessarily be irrational because they are violating "Common Knowledge of Rationality, not rationality" (p. 118). That is, if backwards induction is implausible, as it is in a Centipede game for instance, or if we see that players do not seem to be using backwards induction, as in experimental settings, this is reason to reject Common Knowledge of Rationality as a premise, not to believe that the players themselves are acting irrationally. Gintis argues "Common Knowledge of Rationality is a powerful and often highly implausible assumption concerning the community of mental representations across Bayesian agents" (p. 119).

One implication of this argument is that common knowledge is a kind of social epistemological achievement. It is essential to all kinds of everyday coordination and interaction, but it cannot be assumed to be merely a necessary implication of individual rationality. It must be explained independently. Gintis argues that common knowledge is an irreducible, emergent phenomenon of human interaction. We need, what Gintis call, a "truly social epistemology" (p. xiv). This recognition motivates a move away from the Nash equilibrium as the dominant equilibrium conception towards the correlated equilibrium.

In a correlated equilibrium, players refer to an external correlating mechanism, what Gintis often calls a choreographer, to allow them to coordinate on an equilibrium point. The choreographer acts as an additional player in the game who moves before the other players in the game move. Using a correlated equilibrium transforms the original game into a new game with an initial move by a choreographer.

The basic idea is easy to understand. Consider a group of musicians preforming a piece of music; the presence of a conductor seems superfluous. After all, each musician have a copy of the score; each violinist, flautist, oboist, will play their part, adding up to the piece of music as a whole. Without a conductor, though, each player may begin a section slightly earlier or later than his or her fellows or play some notes a little longer than another person or otherwise deviate from what someone else might be playing. The conductor acts as an external reference point or anchor. The musicians know that everybody else is also organizing their performance around the conductor. Gintis argues that social norms and rules act as a kind of conductor to organize and harmonize our social life.

The need for external choreography becomes clear when we look at the Folk Theorem for repeated games. The folk theorem states that individuals who play repeated games will stumble onto a mutually beneficial equilibrium. Gintis claims that the folk theorem is so impressive because it explains Adam Smith's insight that individuals motivated to achieve their own ends will act in a way that is beneficial to all. The folk theorem explains this insight in an analytically rigorous way. It shows that all we need to explain social cooperation is individual rational agents that interact over a period of time. This model, however, relies on implausibly strong common knowledge requirements and the possibility of perfect punishment for those who do not cooperate. Gintis argue that explaining human behavior requires that we understand humans as being psychologically attracted to respond to correlating mechanisms in the form of social norms. We have a normative disposition to follow certain social norms that is over and above the benefit we get from following those norms.

Gintis shows how norms of property act as a correlating mechanism in the classic Hawk-Dove game. The Hawk-Dove game is an evolutionary game where players in a population represent strategies. In this game there are two strategies, Hawk and Dove. If a Hawk encounters a Dove, the Hawk will fight the Dove and take its territory. If a Dove encounters a Dove, both will engage in a display of aggression and will then split the territory. If a Hawk encounters another Hawk, they will both fight and be injured. The ability to gain territory without injury correlates to reproductive ability where strategies that reproduce constitute a larger and larger share of the population over time. If there is a clear way to determine which player found a piece of territory first, property, there is an evolutionary stable strategy whereby an agent plays Hawk on their own territory and Dove on another's territory. Property acts as a social norm that allows players to correlate their strategies so that they have a greater probability of preserving their own territory while decreasing the need to engage in costly violence. The existence of an external correlation mechanism, in this case, norms of property, enables the development of a stable equilibrium that is mutually beneficial.

There is a wealth of evidence that shows how interactions between human children and non-human animals often mirror the Hawk-Dove game. Property equilibria arise out of the behavioral propensities that we share with many other types of animals. This equilibrium is a combination of a natural physiological propensity to loss-aversion, that is, the desire to keep what we have even at the cost of a good chance of gaining something more, with the need for a strong and reliable external coordinating mechanism. Possession of some territory or object acts as the coordinating mechanism. Once it is clear that someone possesses or owns something we realize that to take it from them will require a fight and we know that they will fight hard to keep it. Locke and Hume were right—there is a natural tendency to see the world as segmented in terms of individual property. The power of Gintis' approach shows how social norms like property serves a social function, but he also locates our motivation for following social norms in our psychological dispositions.

All of this leads to Gintis' most philosophically interesting thesis: that social and political orders are the result of complex emergent social norms that are themselves the product of psychological and biological predispositions. These predispositions are the product of biological and cultural evolution. As he writes, "Complex norms may be taught, learned, and internalized, but individuals must be genetically predisposed to recognize and obey social norms" (p. 248). Even social norms that emerge from a cultural or historical processes are ultimately the result of biological predispositions and propensities to be attracted to and understand those norms. Human beings have, according to Gintis, a biologically based normative predisposition to follow rules even when those rules can be costly to the individuals involved. Without this disposition, we would not be able to correlate effectively even with the existence of a correlating mechanism. This disposition arose through a process of coevolution between genes and culture that produced a human psychology that could both recognize social norms and rules easily, that is disposed to follow the rules, and to punish those who did not follow the rules. We are disposed to do this even at some cost to ourselves.

Despite the obvious elegance and explanatory power of Gintis' approach, there are some obstacles to its acceptance by philosophers. First, the book is full of math and relies on some knowledge in that department. Still, even with limited math skills, if one is not discouraged but, instead tries to follow along closely, Gintis explains almost everything he is showing formally, one will gain not only from the substance of the work, but will also leave the book with a better understanding of the formal techniques that Gintis takes to be basic to a proper understanding of society. There is much in this book that is interesting, everyone will get something out of it even if they do not get everything. Second, Gintis will strike some philosophers as being uninterested in what they may consider the most interesting problems. For instance, in his discussion of Common Knowledge of Rationality, Gintis is not concerned with modeling an epistemically unerring idealized agent. Whatever the interest that project may have for some epistemologists, it would be pointless from the point of view of developing a workable model of human practical rationality. Rather than a defect, however, this is, rightly seen, as strength to his approach. It refocuses our interest on the types of practical problems that animated the great social philosophers like David Hume and Adam Smith. Third, on a similar note, moral or political philosophers will likely respond to this work by claiming that Gintis is engaged in merely descriptive rather than normative investigation.

There is some truth to this claim, but only some. It was Hume, after all, who alerted us to the is-ought gap in the first place, not to show that moral theorizing was independent of psychological and social facts, but to show that moral and political theorizing is intimately enmeshed in the material of common human life. Gintis continues that project with vastly more powerful analytical tools and experimental data. Philosophers who ignore this approach will suffer and may miss out on the exiting potential to solve theoretically intractable problems with new and powerful techniques.